Annex B: Quality Assurance

Contents

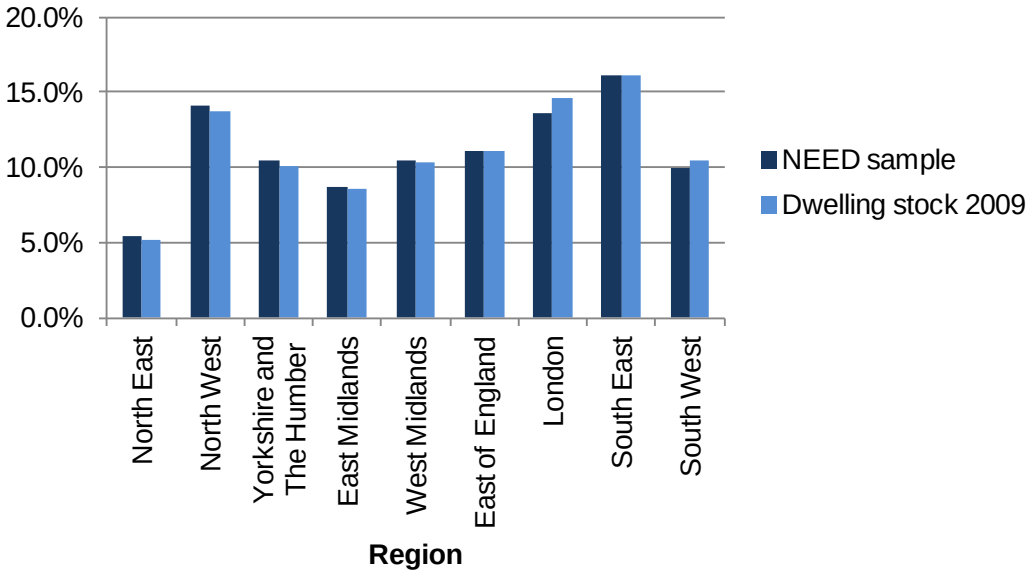
1. Summary and Introduction

The National Energy Efficiency Data-Framework is based on data from a number of sources which are then linked together using the National Land and Property Gazetteer (NLPG) unique property reference number (UPRN). This annex provides information on the quality of each of the datasets used in the domestic NEED analysis and how well they have been matched to form a sample for NEED analysis. For a summary of the data in NEED and how the framework was created see Annex A for results from NEED see the main report and Annexes C and D.

In all cases the match rate between the dataset and the NLPG UPRN was good, the lowest match rate was for the Experian data at 82 per cent. All other domestic datasets had a match rate of at least 94 per cent.

Records for 16 per cent of properties in England were then selected as a sample for analysis. Figure 1.1 shows how the distribution of properties in the NEED sample compares with the Department of Communities and Local Government (DCLG) estimates of dwelling stock in England by region in 2009. It shows that the NEED sample is representative of the population at this level.

Figure 1.1: Comparison of distribution of NEED sample and DCLG dwelling stock estimates



The rest of the annex sets out more details about the quality of the data used in NEED and how the NEED sample compares with the population of English households for each of the variables included in the analysis. Table 1.1 summarises the strengths and weaknesses of each of the main data sources.

Table 1.1: Strengths and weaknesses of data in NEED.

Data sources	Strengths	Weaknesses
Consumption data	• Covers Great Britain. • Good coverage of almost all properties (even post matching). • Data provided by energy suppliers. • Gas data are weather corrected.	• Based on billing data (sometimes estimated). • Gas and electricity years don't cover calendar year (or the same period as each other). • Domestic non-domestic split.
Homes Energy Efficiency Database (HEED)	• Has data for measures installed in homes in the UK including where measure is installed, and date of installation.	• Only covers measures installed through Government schemes; no information on measures installed by households themselves or installed when the property is built. • Matching of (converted) flats not reliable. • Only has a record for 57 per cent of properties in the NEED sample.
Valuation Office Agency (VOA)	• Covers every property in England and Wales. • Excellent coverage– data for each variable used is available for at least 99 per cent of properties in the NEED sample.	• No data for Scotland. • Some data may not be up to date.
Experian	• Data available for each household in the UK. • Best source of data at property level on household characteristics.	• Modelled data with variable accuracy at property level.

The quality and coverage of the data are good but any interpretation of results should be considered in the context of the strengths and weaknesses of each of the data sources used.

1. Matched Sample

As described in Annex A, a sample was used for all the domestic NEED analysis set out in this report. This section provides further information about the sample and how it compares with the distribution of data at the national level. Further information about the quality of the datasets and how representative the sample is are provided in sections three to six of this annex.

Address information in each dataset was matched to the address information on the National Land and Property Gazetteer (NLPG)[1] in order to assign a unique property reference number (UPRN) to each record. The table below shows what proportion of records in each dataset could be matched to a UPRN.

Table 2.1: Match rates (sub-building[2] match rates in brackets).

Data source	Match rate
Electricity consumption	94% (87%)
Gas consumption	97% (93%)
HEED	99% (98%)
Experian	82% (69%)
VOA property attribute data[3]	100%

The match rates set out in the table are calculated based on the number of records for each relevant source; not the number of UPRNs in England and Wales. For example, DECC only leased Experian data for a representative sample of about 3 million properties; the match rate shows how many of these 3 million records could be matched to the NLPG. The figure quoted for HEED excludes flats which were excluded from the analysis of impacts of energy efficiency measures due to difficulties with matching at sub-building level. The electricity and gas consumption figures quoted cover domestic and non-domestic properties.

A sample of the data was then used for analysis. The sample was made up of 3.6 million records, or 16 per cent of properties in England. The sample was selected from the VOA dataset from those records which had a valid UPRN. Therefore the VOA dataset has a match rate of 100 per cent. The UPRN was then used to link data from the other sources to create the analysis sample.

For a number of reasons, including but not limited to matching, records in the NEED sample do not always have information for each variable. The tables below show how well populated each of the variables used in NEED were in the NEED analysis sample.

Table 2.2 shows the number and proportion of records which had gas and electricity consumption in each year from 2004 to 2010.

---

[1] See Annex A for more information about the NLPG.
[2] A sub-building is a separate property within the same building. Such as a flat within a converted property or an individual shop within a shopping centre.
[3] The match rate for VOA data is 100 per cent as only VOA records that could be matched to the NLPG were included in the sample.

Table 2.2: Records in NEED sample with a consumption value 2004 to 2010[4]

	Gas		Electricity	
	N	%	N	%
2004	2,904,560	81%	3,433,710	95%
2005	2,995,180	83%	3,417,740	95%
2006	3,023,730	84%	3,512,750	98%
2007	3,050,510	85%	3,510,950	98%
2008	3,050,510	85%	3,513,250	98%
2009	3,059,450	85%	3,536,480	98%
2010	3,057,990	85%	3,550,740	99%

It shows that between 80 and 85 per cent of records had a gas consumption value in the matched dataset, and between 95 and 99 per cent of records had a value for electricity. A small number of missing records are a result of consumption records which could not be matched to the UPRN. The additional differences are primarily as a result of properties that do not have a gas connection or do not use gas. The number of populated records is lower in earlier years as the matching was carried out on address information associated with 2009 meter addresses.

It should be noted that this table includes all records with any gas consumption. However, a number of the records which have been included in the dataset do not have valid consumption values, more information on this is included in Section 2 of this annex.

Table 2.3 shows the equivalent information for the Homes Energy Efficiency Database (HEED).

Table 2.3: Records in NEED sample with a HEED value

	Count	
	N	%
HEED Record	2,044,820	57%
Cavity wall insulation	640,890	18%
Loft insulation	536,410	15%
Heating measure	355,320	10%
Solid wall insulation	1,670	0%

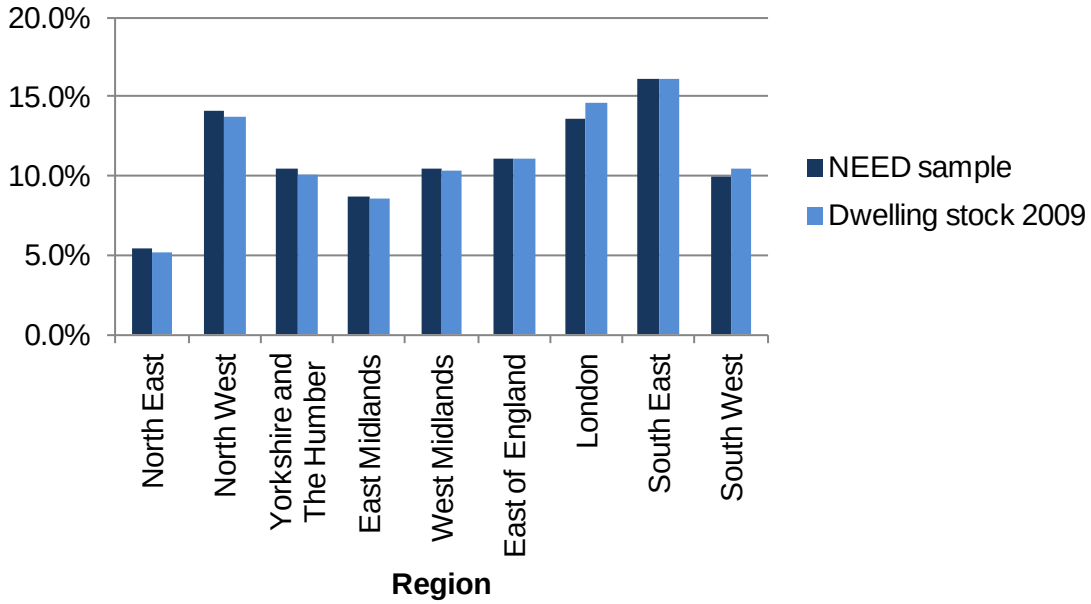
The table shows that a HEED record exists for 57 per cent of properties in the NEED sample. However, in many cases these records may not include any information relevant to the NEED analysis. The figures for cavity wall, solid wall and loft insulation and for boilers are also shown in the table. As described in more detail in Section 3, this only includes information on measures recorded in HEED (primarily retro-fit insulation installed through Government schemes). It does not cover all measures in all properties in England. All records with a measure installed had an associated installation date in the dataset. However, the dates were not valid in all cases.

[4] The number of records quoted here includes flats which are excluded from the analysis of the impact of energy efficiency measures in section 5 and Annex D (they are included in the domestic consumption analysis).

As the dataset was based on records that existed in the VOA dataset there was no loss of VOA property attribute information due to matching. For all VOA variables used in the NEED analysis 99 per cent or more of records had a valid value. Section 4 of this annex gives more detail.

For the Experian data, data were available for 81 per cent of records. For geographic information, data, including region, local authority and output area classification, were available for all records. Figure 2.1 shows how the distribution of records by region in the NEED sample compares with DCLG's published estimates of dwelling stock by region.

Figure 2.1: Comparison of distribution of NEED sample and DCLG dwelling stock estimates[5]



The chart shows that the distribution of the NEED sample accurately reflects the distribution of the total dwelling stock in England at region level. Further details of how the NEED sample reflects the distribution of the housing stock in other respects is included in the relevant sections below.

[5] DCLG Table 109: http://www.communities.gov.uk/housing/housingresearch/housingstatistics/housingstatisticsby/stockincludingvacants/livetables/

2. Consumption Data

3.1 Introduction

UK Government has collected and published energy consumption data within the Digest of UK energy Statistics since 1948[6] and time series back to 1970 on how energy has been used is published in Energy Consumption in the UK[7]. However, data at individual meter point, as used in NEED, was first obtained in 2004 in order to produce local areas estimates of consumption – work that was awarded a Royal Statistical Society Award for innovation in 2010. These data cover consumption of gas and electricity for all homes and businesses within England, Scotland and Wales. There is no property level data available for other fuels which may be being used to heat homes, such as oil or coal. The electricity and gas data are from energy suppliers administrative systems and cover around 30 million electricity meters and 25 million gas meters. The consumption data are published on the DECC website down to Lower Level Super Output Area (groups of approximately 400 homes)[8].

Data for 2004 consumption are referenced in some places in the report, these data were experimental, and should therefore be used with caution. It is recommended that 2005 data are used for baselines.

This section provides more detail on the electricity and gas consumption data used in NEED, including; a summary of where the data come from; a description of the data; and how the data are cleaned before being used in NEED. It also outlines rationale for any differences in the approach used in NEED compared with the approach used in DECC's sub-national consumption publication and shows how the data used in NEED compare with consumption data reported by other sources.

3.2 Gas consumption data

Data collection

DECC obtain annualised consumption estimates for all gas meters in Great Britain. The majority come from xoserve, the company responsible for the collation and aggregation of gas consumption, with a further (approximately) one million provided by the independent gas transporters. DECC are provided with annualised estimates of consumption for all the MPRN's (meter point reference numbers) in Great Britain based on an Annual Quantity (AQ). An AQ is an estimate of annualised consumption using consumption recorded between two meter readings at least six months apart. The estimate is then adjusted to reflect a 17 year weather correction factor. The AQ for each MPRN represents consumption relating to the gas year – the period covering 1 October through to the following 30 September.

---

[6] http://www.decc.gov.uk/en/content/cms/statistics/publications/dukes/dukes.aspx
[7] http://www.decc.gov.uk/en/content/cms/statistics/publications/ecuk/ecuk.aspx
[8] More detailed information about how these data are collected and compiled for DECC's sub-national publication is available on the DECC website: http://www.decc.gov.uk/assets/decc/Statistics/regional/1087-guidance-note-regional-energy-data.pdf.

The data are provided with permission from the owners of the local distribution zones (LDZ) network (i.e. the four major gas transporters in Great Britain – National Grid, Scotia, Wales and West Utilities and Northern Gas Networks) and agreement by the gas suppliers.

The gas data has no reliable domestic and industrial/commercial flag to enable an accurate split between these sectors. The gas industry use a crude cut off of 73,200 kWh, with customers using less than this assumed to be domestic. This cut off is therefore also used in DECC's published sub-national consumption publication. This means that in the sub-national estimates, there are a significant number of businesses (estimated to be around 2 million) misallocated. This is an issue which DECC are looking to resolve, but does not impact on data in NEED. NEED uses the allocation of property for council tax and non-domestic rates to define which customers are domestic and which are non-domestic. There are some limitations to this approach, particularly for the non-domestic section which is covered in Section 6 of the NEED report. However, it is believed to be considerably more accurate than the crude approach used by the gas industry.

Coverage

The gas data exclude properties in Northern Ireland, due to the market structure. In addition, a considerable amount of consumption relating to power stations and some very large industrial consumers is not included in the data.

The data represent gas transported through the national distribution system and gas that passes through the National Transmission System into other independently owned local distribution systems. However, the data exclude any gas passing through other transmission and distribution systems such as those owned by North Sea producers. It also excludes large loads fed directly from the National Transmission System (such as certain power stations and large industrial consumers).

The data do include the 2,500 gas consumers whose consumptions are recorded on a daily basis who are known as Daily Metered (DM) customers. The non domestic section of the NEED report (Section 6) provides more information on quality of data for the non-domestic sector. Further information on the quality of data in the domestic sector is set out below.

Summary of data

The gas dataset received from the gas transporters is matched to the NLPG via the address information and assigned a Unique Property Reference Number (UPRN) before it can be used in NEED. The table below summarises the data provided by GB Group once the original dataset had been matched to the address identifier, but before it was matched to other datasets. Note the data in the table below includes domestic and non-domestic data and shows mean and median consumption before any cleaning and validation of the dataset.

Table 3.1: Mean and median gas consumption (kWh) pre-matching to other datasets and number of meters contributing[9]

	2004	2005	2006	2007	2008	2009	2010
N	21,243,000	21,992,000	22,263,000	22,575,000	22,651,000	22,873,000	23,003,000
Mean	28,200	27,300	26,000	24,900	23,700	21,600	23,500
Median	18,300	17,900	17,000	16,300	15,600	14,000	13,800

These data were then used when the analysis file was created (see Annex A for more details on how the analysis sample was selected). Table 3.2 below shows the distribution of the data in the analysis file before any cleansing of the data had taken place. Because of the nature of the sample selection, the data in the table should all refer to domestic consumers. However, there are some cases where the consumption for a property is above the threshold considered valid for domestic consumers.

Table 3.2: Mean, median (kWh) and number of missing records

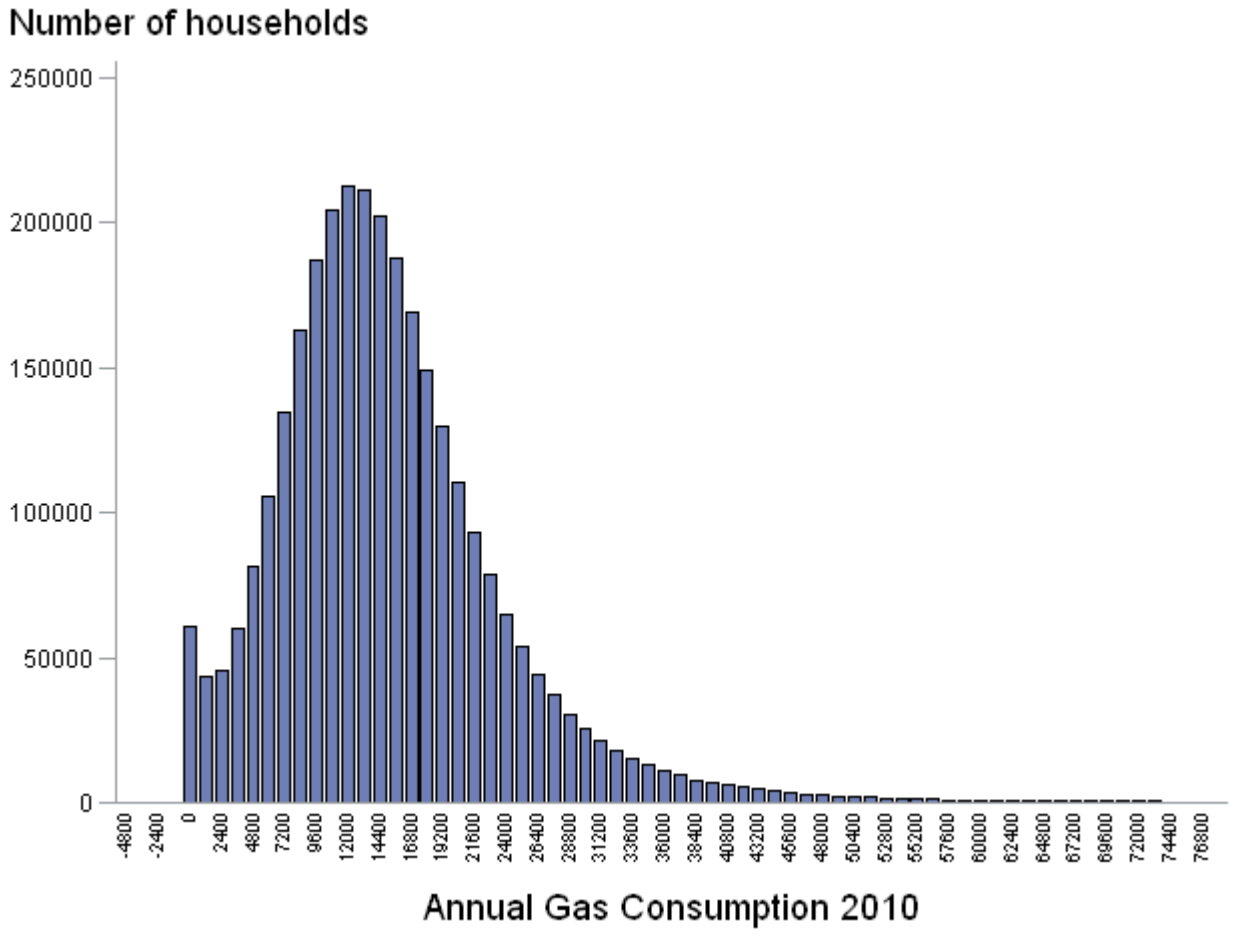
	2004	2005	2006	2007	2008	2009	2010
N	2,904,560	2,995,180	3,023,730	3,050,510	3,050,510	3,059,450	3,057,990
Mean	20,200	19,600	18,800	18,100	17,300	15,700	15,500
Median	18,200	17,800	17,000	16,300	15,600	14,100	13,900
Missing (N)	694,870	604,250	575,700	548,920	548,920	539,980	541,440
Missing (%)	19.3%	16.8%	16.0%	15.3%	15.3%	15.0%	15.0%

Records without gas consumption data (identified in the table as missing) have two main causes; properties that are not on the gas network, or do not use gas – in 2010 approximately 83% of homes in Great Britain had gas meters[10] – and records which could not be matched to a valid UPRN. Figure 3.1 shows the distribution of the data in the sample.

[9] These data differ from the data reported in the sub-national consumption statistics. The main reasons for these differences are the validation that has been carried out on the sub-national consumption statistics before publication. The data in the table above include some records which are not included in the publication because of quality. It is a small number of records but they are having a significant impact on the mean consumption values shown in the table. For example, the difference between mean consumption in 2008, 2009 and 2010 primarily reflects the inclusion of a small number of erroneous records rather than a general population change.

[10] Based on DECC's sub-national statistics.

Figure 3.1: Distribution of gas consumption data pre-cleansing.



The frequency chart above shows all the records in the dataset between 0 and 73,200 kWh (the industry cut off for domestic consumption). In addition, 0.3 per cent of gas values in the dataset were greater than 73,200 kWh.

Further work was done to look at the gas consumption data and assess which records are valid and should be included in the analysis. The next section outlines the assessment of the data and why decisions on inclusion of data have been made.

Data validation

Consistent with the approach taken for sub-national statistics publications, the NEED analysis started by excluding any records with consumption greater than 73,200 kWh as it is assumed they are not domestic. However, because of the nature of the analysis undertaken in NEED further cleansing and validation was undertaken. This means that consumption figures in NEED are not the same as those in the sub-national consumption publication, but are very similar. For example, unlike the sub-national consumption statistics, all negative meter readings are also excluded[11].

Table 3.3 shows the percentiles for the data in the analytical sample with a consumption between 1 and 73,200 kWh.

[11] As data are based on billed consumption, it is possible that a negative reading is valid if an estimated reading provided in a previous year had been too high. However, these reading are not considered valid in NEED.

Table 3.3: Distribution of gas consumption data in NEED sample pre-cleansing (kWh)

Annual consumption (kWh)	2004	2005	2006	2007	2008	2009	2010
1st Percentile	236	251	151	133	69	32	49
5th Percentile	4,119	4,287	3,797	3,645	3,210	2,989	3,054
Lower Quartile	12,487	12,253	11,597	11,109	10,501	9,499	9,447
Median[12]	18,149	17,727	16,903	16,219	15,551	14,045	13,883
Upper Quartile	24,567	23,938	22,986	22,319	21,475	19,490	19,204
95th Percentile	38,367	37,682	36,564	35,714	34,569	31,534	30,979
99th Percentile	54,631	54,103	53,172	52,325	50,845	47,199	46,185

From Table 3.3. and Figure 3.1 it is clear that the majority of gas consumption data is below 50,000 kWh. In order to avoid the relatively small number of properties with consumption over 50,000 kWh having a disproportionate impact on the analysis in NEED these have been excluded. This should reduce the likelihood of including non-domestic properties or domestic properties with invalid consumption in the analysis. All analysis excludes records with a gas consumption of more than 50,000 kWh in the year being considered.

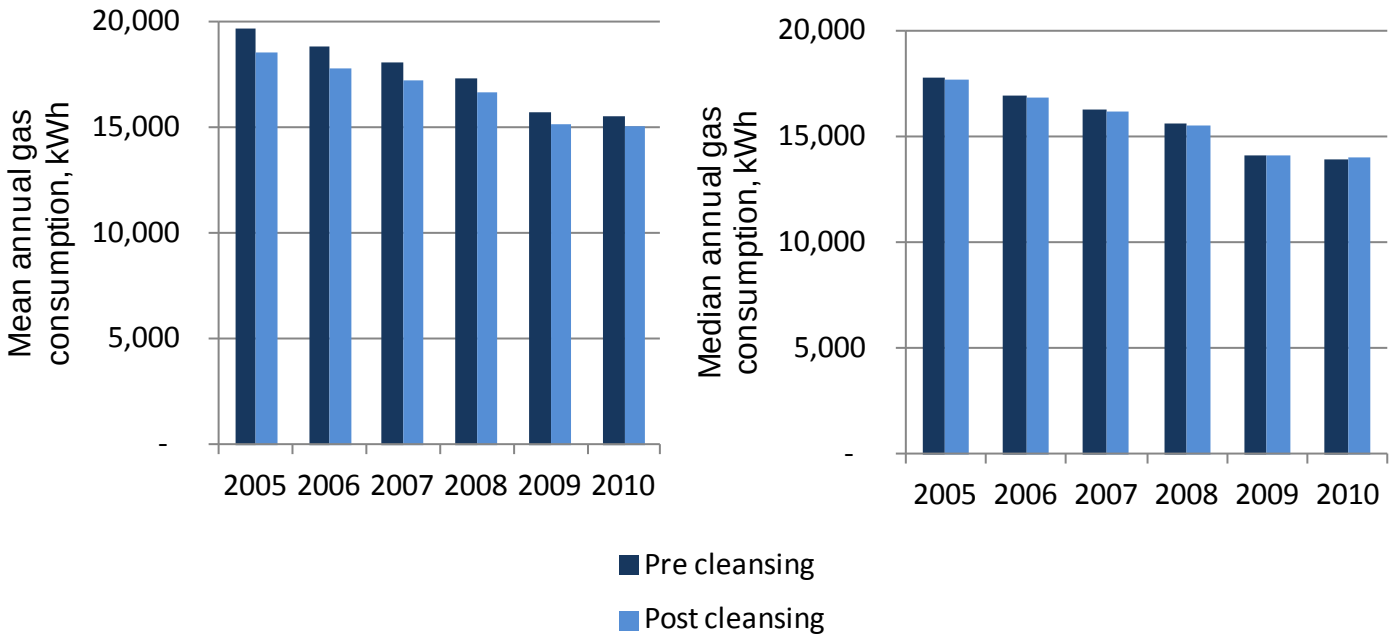
At the lower end of the distribution, there are a cluster of values around 1 kWh to 100 kWh. These have also been excluded from all analysis, as they are likely to be households with gas supplies which are not used (or new build properties which are not yet occupied).

In addition, suspected estimated values have been excluded from the data before analysis was undertaken. These take two forms. For any given year, if a household has a consumption value identical to the previous year it is assumed to be an estimate. There are also a small number of values which are suspected to be estimated readings used by suppliers. These were assumed on the basis of values that appear in the data more often than would be expected given the frequency of similar consumption values. For example, if there were five times more properties with consumption of 10,500 than there were with consumption of 10,499 kWh or 10,501 kWh then all consumption values of 10,500 kWh would be excluded from the analysis.

Figure 3.2 shows the mean and median consumption for data in the NEED sample before and after these filters had been applied. The median values are very similar in both cases, however the mean consumption is lower for the cleansed dataset. This is because of the elimination of a relatively small number of records with a high consumption which were having a disproportionate influence on the mean.

[12] This differs from the median consumption shown in Table 3.2 as data in Table 3.3 are limited to those in the NEED sample with a consumption between 1 to 73,200 kWh in the specified year.

Figure 3.2: Mean and Median consumption before and after data cleansing has been applied



Filters for analysis of impacts of energy efficiency measures

For analysis relating to gas heating, a higher cut off, of 2,500 kWh, has been put in place. This cuts out approximately three per cent of the sample of valid gas consumption, but helps to ensure that the properties that are included in the analysis of the impact of energy efficiency measures are using gas as their main heating fuel. To help inform the relevant cut offs for gas heating the EHS modelled data matched with the NEED consumption data was considered. More information about this pilot work can be found in Section 6.1 of the Annual Report on Fuel Poverty Statistics 2012 here: http://www.decc.gov.uk/assets/decc/11/stats/fuel-poverty/5270-annual-report-fuel-poverty-stats-2012.pdf. It showed that almost all households in the matched dataset that use gas as their main heating fuel had annual gas consumption (both modelled and actual) of 2,500 kWh or more.

In addition to the validation above further filters were applied for the impact of measures work. These included restricting the change in consumption for the period prior to a measure being installed and the year after the measure was installed. The maximum decrease in consumption over the period was 80 per cent and maximum increase in consumption was set at 50 per cent. The purpose of this is to eliminate households which have extreme changes in consumption as these are likely to be due to a change in circumstances or occupier rather than as a result of the measure which had been installed. All filters used in impact of measures work have been applied in the same way to the households which had received measures and to households which make up the comparator group[13].

[13] Flats have also been excluded from the analysis of impact of energy efficiency measures work due to difficulties matching HEED data to the correct property within a building.

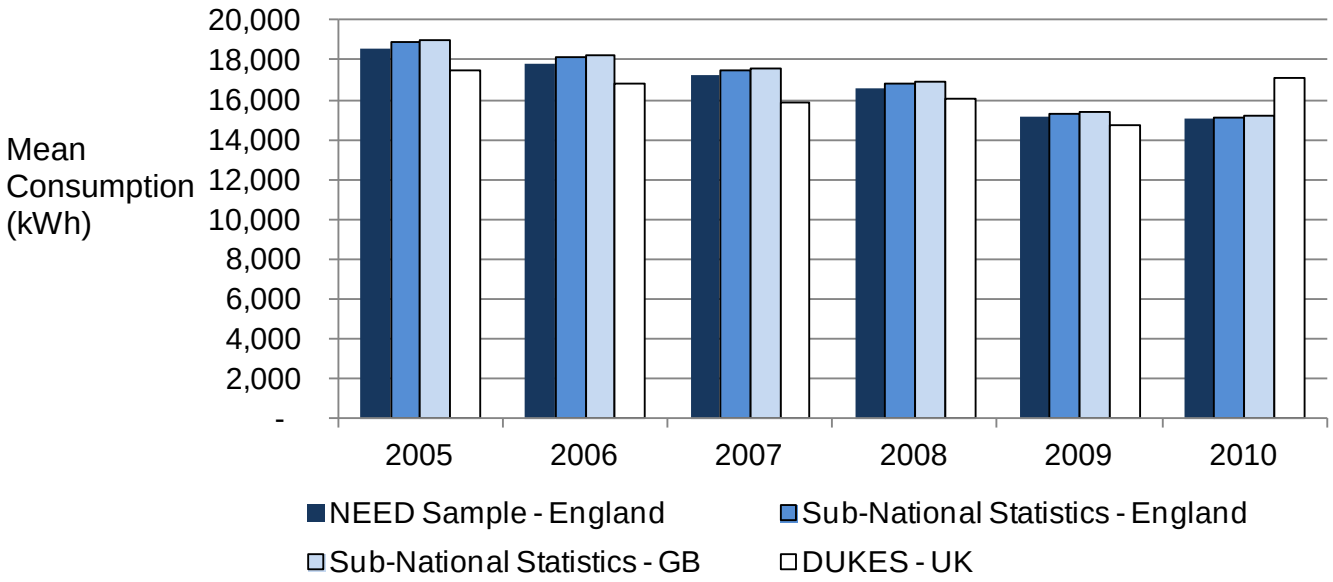
Table 3.4: Summary of cleansing applied to gas consumption data used for analysis in NEED.

Analysis of consumption	Analysis of impact of energy efficiency measures
• Consumption between 100 kWh and 50,000 kWh.	• Consumption between 2,500 kWh and 50,000 kWh - in year measure installed as well as year prior to and after measure installed.
• Removed suspected estimated values.	• Removed suspected estimated values.
	• Maximum decrease in gas consumption of 80% over period analysed[14].
	• Maximum increase in consumption of 50% over period analysed[15].

Comparison with other sources

To check that the sample used for analysis is consistent with the other estimates of domestic consumption published by DECC - and therefore increase confidence in use of the data – the NEED analysis sample after cleansing has been compared with the data from the Digest of UK Energy Statistics (DUKES)[16] and the data from the sub-national consumption statistics published by DECC.

Figure 3.3: Comparison of estimates of mean gas consumption per household from different sources

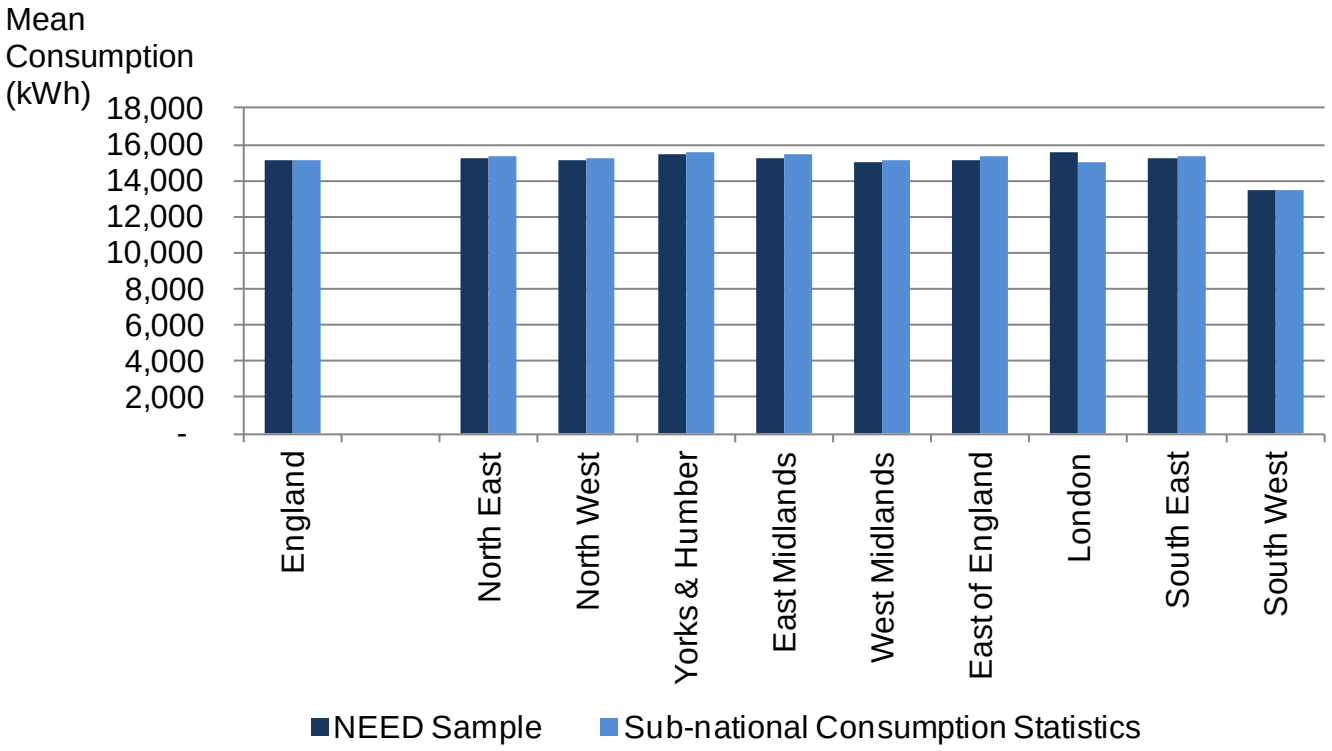


[14] E.g. for measures installed in 2009, any property with a decrease in consumption of more than 80 per cent between 2008 and 2010 is excluded from the analysis.
[15] E.g. for measures installed in 2009, any property with an increase in consumption of more than 50 per cent between 2008 and 2010 is excluded from the analysis.
[16] http://www.decc.gov.uk/en/content/cms/statistics/publications/dukes/dukes.aspx

Figure 3.3 shows that the mean consumption is very similar for the published sub-national statistics and for the NEED sample. It is broadly similar to the DUKES estimates, however there is a lot more variation in the DUKES data between years. The primary reason for this variation is the difference between sources. DUKES data is not weather corrected, while the NEED and sub-national statistics data are weather corrected. Therefore in a cold winter, the DUKES data shows higher mean consumption, while in a year that has a warmer winter the DUKES mean consumption is lower than that given by NEED and the sub-national statistics[17]. There will also be some variation in the difference between the DUKES and sub-national Statistics or NEED estimates because of the method of collection. While the later two are based on the same source data and built up from households data, DUKES is based on aggregate data. Some of the variation is also due to the different geographic coverage of the different datasets. The sub-national statistics estimates for England and Great Britain are included to give an indication of the impact this is likely to be having.

Figure 3.4 below shows how the data compare for the sub-national statistics and the NEED sample at region level.

Figure 3.4: Mean gas consumption per household 2010 by region, sub-national consumption estimates compared with NEED analysis sample.



The chart shows that the cleansed NEED sample and the sub-national consumption published data have a very similar mean consumption in all regions. The biggest difference seen is for London where the sub-national consumption estimates are four per cent lower than the mean from the NEED sample (primarily because flats are underrepresented in the NEED sample). In all other regions the NEED sample has a mean consumption which is lower than the sub-national mean (approximately one per cent or less in all cases). This is because of the 50,000

[17] The published DUKES data gives a total domestic consumption figure for the UK which has been converted into a mean consumption based on the number of dwellings in the UK and the proportion of households in GB with gas meters.

kWh maximum consumption used for the NEED analysis. Although only a small number of records are eliminated, these records have a disproportionate impact on the mean.

3.3 Electricity consumption data

Data collection

Data are collected with the full co-operation of the electricity industry. Annualised consumption data are generated by the data aggregators, agents of the electricity suppliers, who collate/aggregate electricity consumption levels for each customer meter or MPAN (meter point administration number). In addition to this, address information for each meter is obtained from the Gemserv meter address file.

The electricity consumption data are generated for both non half hourly (NHH) meters (domestic and small/medium commercial/industrial customers) and for half hourly (HH) meters (larger commercial/industrial customers). There are around 29 million NHH meters and 113,000 HH meters in Great Britain. For the NHH data, annualised estimates are based on either an annualised advance (AA) or estimated annual consumption (EAC). The AA is an estimate of annualised consumption based on consumption recorded between two meter readings. In comparison an EAC is used where two meter readings are not available and an estimate of annualised consumption is produced by the energy company using historical information and the profile information relating to the meter. These data provide a good approximation of annualised consumption, but do not cover exactly the calendar year. For example, 2010 annualised consumption estimates cover the 365 days up to 30 January 2011. For the half hourly meter consumption estimates, data aggregators are asked to produce a simple report for each MPAN for the relevant calendar year.

DECC publish estimates of consumption with domestic/non-domestic splits, with aggregate and average consumption figures provided for each local authority and region. The domestic consumption is based on NHH meters with profiles 1 and 2 (these are the standard domestic and economy 7 type tariffs respectively). Non-domestic consumption is based on NHH meters with profiles 3 to 8 and all HH meters (and any nominally domestic meters with consumption of more than 100,000 kWh in a year or meters with consumption between 50,000 and 100,000 kWh with address information which suggests non-domestic use). However, it should be noted that these assumptions differ from those used in NEED, where the use of the data means it is more appropriate to use a slightly different approach to ensuring a property is domestic and has valid consumption. This is described in more detail in data validation section below.

Coverage

These data cover all of Great Britain. Data for Northern Ireland are currently excluded from the dataset (though work is on going to produce data for Northern Ireland and experimental data has been published in Energy Trends[18]). Some very large industrial consumers with connection to high voltage lines of the transmission system are also excluded. These consumers are classified as CVA or Central Volume Allocation users, who have different arrangements with their electricity suppliers, compared to NHH and HH meter customers. CVA generally accounts for around 2% of electricity sales.

[18] http://www.decc.gov.uk/assets/decc/11/stats/publications/energy-trends/3917-trends-dec-2011.pdf

Summary of data

Before it can be used in NEED, the electricity data are matched to the National Land and Property Gazetteer (NLPG) via the MPAN address information and assigned a Unique Property Reference Number (UPRN). The table below summarises the data provided by GB Group once the original dataset had been matched to the address identifier, but before it was matched to other datasets used to form NEED. Note the data in the table below includes meters of all profiles (domestic and non-domestic data) and shows mean and median consumption before any cleansing and validation of the dataset.

Table 3.5: Mean and median electricity consumption (kWh) pre-matching to other datasets and number of meters contributing[19]

	2004	2005	2006	2007	2008	2009	2010
N	28,128,510	28,163,940	29,014,910	28,959,200	29,089,930	29,361,940	29,589,170
Mean	6,300	6,300	10,600	10,300	10,100	9,800	8,700
Median	3,800	3,700	3,700	3,600	3,400	3,400	3,400

These data were then used when the analysis file was created (see Annex A for more details on how the analysis sample was selected). The table below shows the distribution of the data in the analysis file before any cleansing of the data had taken place. Because of the nature of the sample selection, the data in the table should all refer to domestic consumers. However, further cleansing, as described below, was carried out to ensure the none of the anomalous consumption values were included in the final analysis.

Table 3.6: Mean, median (kWh) and number of missing records of NEED sample pre-cleansing

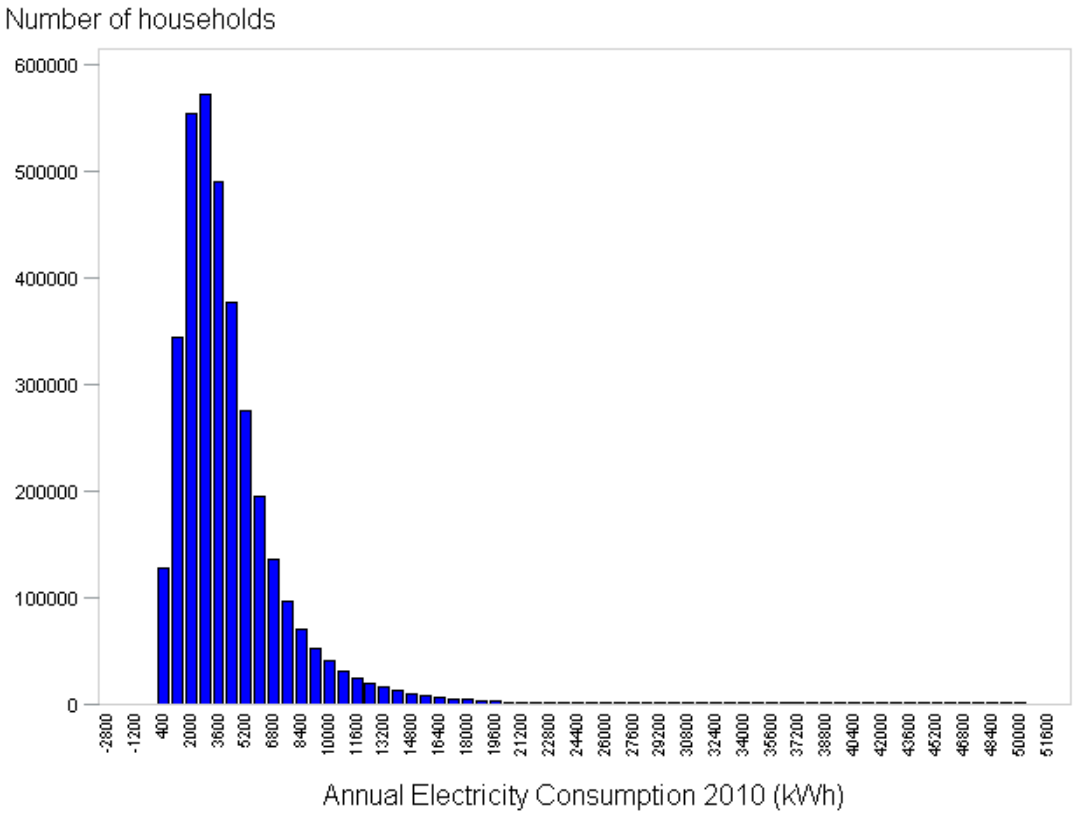
	2004	2005	2006	2007	2008	2009	2010
N	3,433,710	3,417,740	3,512,750	3,510,950	3,513,250	3,536,480	3,550,740
Mean	4,800	4,800	4,700	4,600	4,500	4,400	4,400
Median	3,700	3,700	3,700	3,600	3,400	3,400	3,400
Missing (N)	165,720	181,690	86,680	88,480	86,170	62,950	48,690
Missing (%)	4.6%	5.0%	2.4%	2.5%	2.4%	1.7%	1.4%

Missing records are as a result of properties which were not assigned a valid UPRN and therefore could not be matched to other data in the sample. The matching was carried out on 2009 address information and therefore the match rates are best for 2009 and 2010.

Figure 3.5 shows the distribution of electricity consumption data for the NEED sample. The data included in this figure should all relate to domestic properties as they are based on data in the VOA dataset. Only records with consumption between 1 and 50,000 kWh have been included. In addition, 0.1 per cent of records in the NEED sample had a consumption value of more than 50,000 kWh in 2010.

[19] These data differ from the data reported in the sub-national consumption statistics. The main reasons for these differences are the validation that has been carried out on the sub-national consumption statistics before publication. The data in the table include some records which are not included in the publication because of quality. It is a small number of records but they are having a significant impact on the mean consumption values shown in the table. For example the mean consumption in 2004 and 2005 is not typical of the population, but results from a single erroneous record.

Figure 3.5: Distribution of electricity consumption data pre-cleansing



Further work was undertaken to look at the electricity consumption data and assess which records are valid and should be included in the analysis. The next section outlines the assessment of the data and why decision on inclusion of data have been made.

Data validation

It is assumed that any consumption of over 50,000 kWh is not domestic. However, because of the nature of the analysis undertaken in NEED further cleansing and validation was undertaken to decide on what should be considered valid data for this analysis. This means that consumption figures in NEED are not the same as those in the sub-national consumption publication, but are very similar. For example, unlike the sub-national consumption statistics, all negative meter readings are also excluded[20].

Table 3.7 shows the percentiles for the data in the analytical sample with a consumption between 1 and 50,000 kWh.

Table 3.7: Distribution of 2010 electricity consumption data in NEED sample pre-cleansing (kWh)

	1st Percentile	5th Percentile	Lower Quartile	Median	Upper Quartile	95th Percentile	99th Percentile
2010	217	975	2,193	3,442	5,256	10,425	17,858

[20] As data are based on billed consumption, it is possible that a negative reading is valid if an estimated reading provided in a previous year had been too high. However, these reading are not considered valid in NEED.

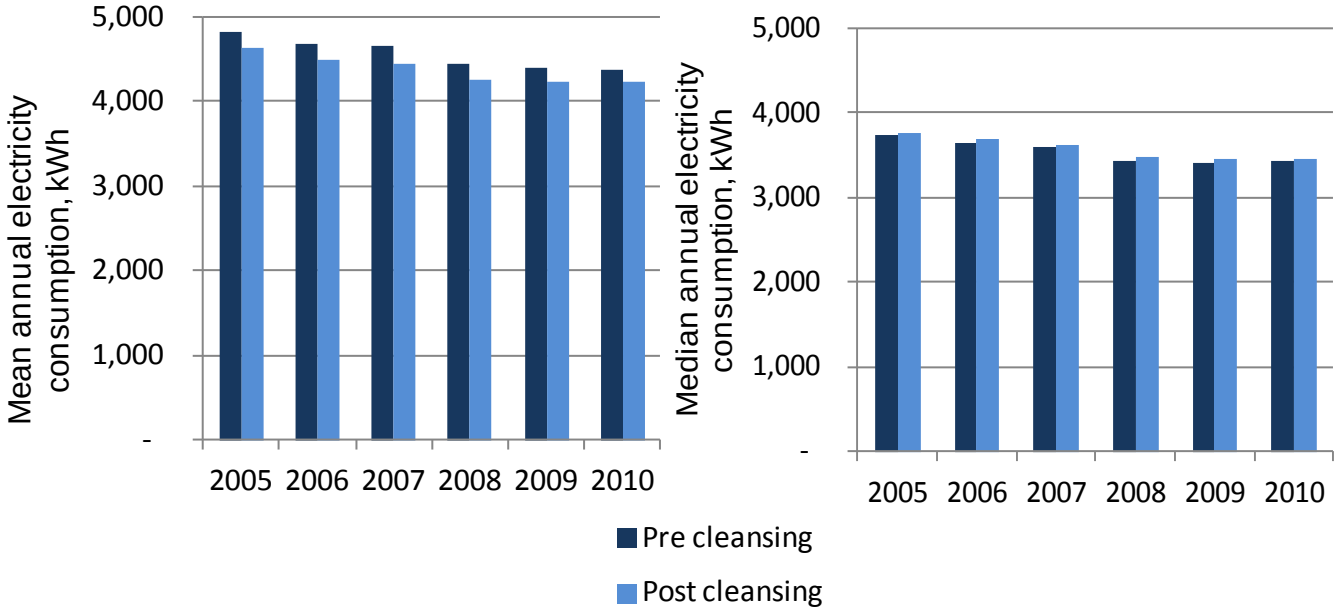
From Table 3.7 and Figure 3.5 it is clear that the majority of electricity consumption data is below 25,000 kWh. In order to avoid the relatively small number of properties with consumption over 25,000 kWh having a disproportionate impact on the analysis in NEED these have been excluded. This should reduce the likelihood of including non-domestic properties or domestic properties with invalid consumption in the analysis.

At the lower end of the distribution, there are a cluster of values around 1 kWh to 100 kWh. These have also been excluded from all analysis, as they are likely to be households with electricity supplies which are not used (or new build properties which are not yet occupied).

In addition, suspected estimated values have been excluded from the data before analysis was undertaken. These take two forms. For any given year, if a household has a consumption value identical to the previous year it is assumed to be an estimate. There are also a small number of values which are suspected to be estimated readings used by suppliers. These were assumed on the basis of values that appear in the data more often than would be expected given the frequency of similar consumption values.

Figure 3.6 shows the mean and median consumption for data in the NEED sample before and after these filters had been applied. The median values are very similar in both cases, however the mean consumption is lower for the cleansed dataset. This is because of the elimination of a relatively small number of records with a high consumption which were having a disproportionate influence on the mean.

Figure 3.6: Mean and Median consumption before and after data cleansing has been applied



Unlike the gas consumption data, electricity consumption data has not been used in the analysis of the impact of energy efficiency measures, so there is no separate cut off for used households which use electric heating. However, there is a big range in electricity consumption, due to the differences in households which use electricity for heating and

households which do not. Comparison of the meter point consumption data with the EHS modelled consumption[21] suggests that households which do not use electricity for heating their homes tend to have a consumption of 10,000 kWh or below. The cut off for electricity for analysis in NEED will be reviewed again prior to future publications and decided based on the analysis being undertaken.

Comparison with other sources

To check that the sample used for analysis is consistent with the other estimates of domestic consumption published by DECC - and therefore increase confidence in use of the data – the NEED analysis sample after cleansing has been compared with the data from the Digest of UK Energy Statistics (DUKES)[22] and the data from the sub-national consumption statistics published by DECC.

Figure 3.7: Comparison of estimates of mean electricity consumption per household from different sources

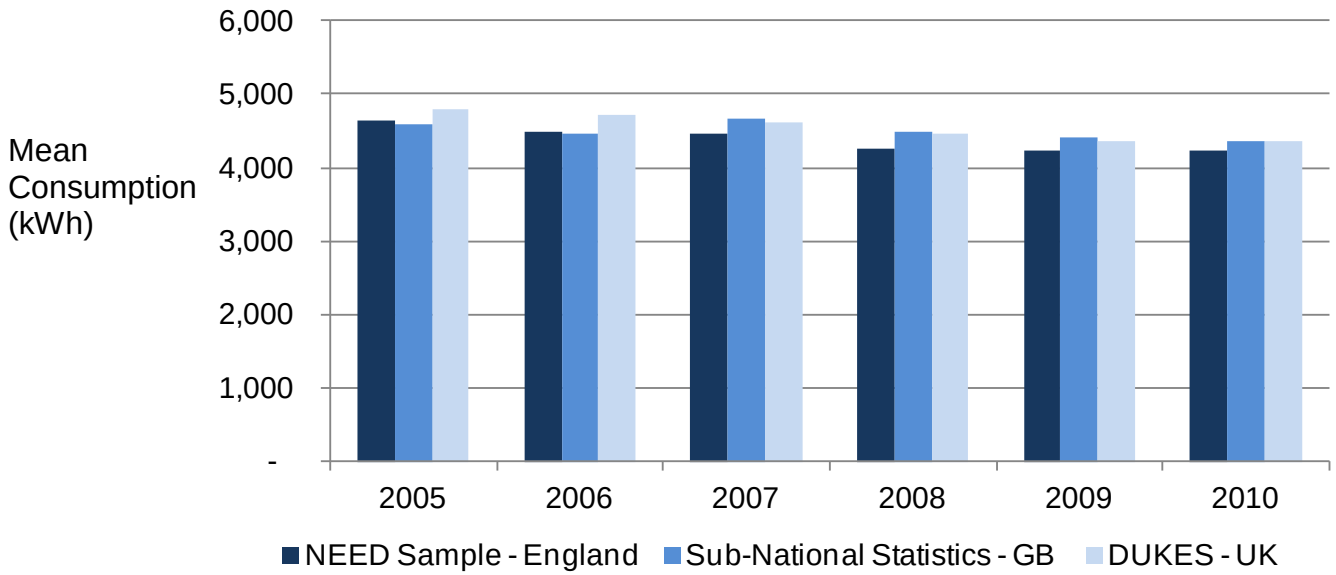
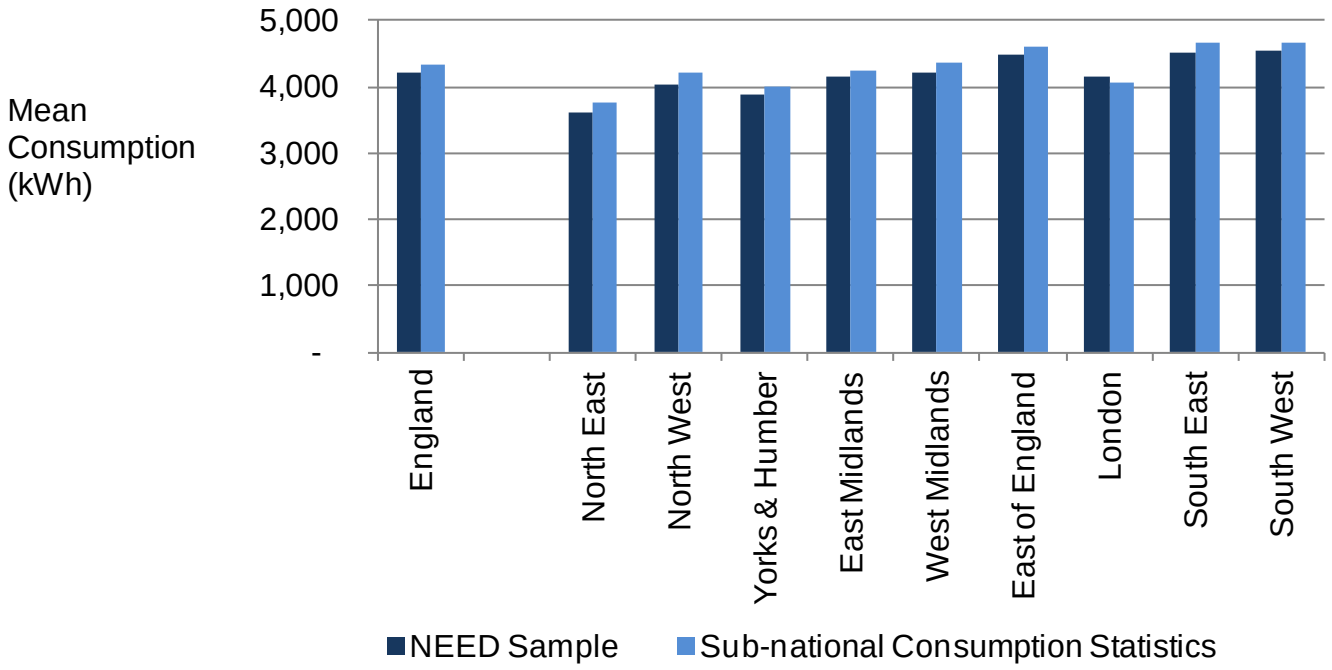


Figure 3.7[23] shows that the mean consumption is very similar for all three sources. The primary reason for this variation is the difference between sources. There will also be some variation in the difference between the DUKES and sub-national statistics or NEED estimates because of the method of collection. While the later two are based on the same source data and built up from households data, DUKES is based on aggregate data. Some of the variation is also due to the different geographic coverage of the different datasets.

Figure 3.8 below shows how the data compare for the sub-national statistics and the NEED sample at region level.

[21] Based on the pilot matching described in Section 6.1 of the Annual Report on Fuel Poverty Statistics 2012 here: http://www.decc.gov.uk/assets/decc/11/stats/fuel-poverty/5270-annual-report-fuel-poverty-stats-2012.pdf.
[22] http://www.decc.gov.uk/en/content/cms/statistics/publications/dukes/dukes.aspx
[23] The published DUKES data gives a total domestic consumption figure for the UK which has been converted into a mean consumption based on the number of dwellings in the UK. The sub-national consumption data is based on consumption per consumer in 2005 and 2006 and sales per household from 2007.

Figure 3.8: Mean electricity consumption per household 2010 by region, sub-national Consumption estimates compared with NEED analysis sample.



The chart shows that the cleansed NEED sample and the sub-national consumption published data have similar mean consumption in all regions. The biggest difference is in the North West where the mean consumption for the sub-national estimates is nearly five per cent higher than the mean for the NEED sample. In all regions except London, the sub-national mean is higher than the mean based on data in the NEED sample. This is because of the lower cut off for valid consumption used in the NEED sample. Although only a small number of records are eliminated, these records have a disproportionate impact on the mean. In London the NEED mean is higher than the sub-national mean. This is because of the poorer quality matching of flats which generally have lower consumption than other properties.

3.4 Conclusion

The consumption data are a very rich source of data which form the core of NEED. The analysis above sets out the further work that has been done to ensure these data are of suitable quality. Table 3.8 summarises how the data used in NEED differs from the sub-national consumption data as a result of the cleansing and validation carried out on the data.

Table 3.8: Differences in consumption data

NEED data	DECC's Sub-National Consumption Estimates
• Property must be included as a domestic property on the Valuation Office Agency property attribute dataset to be included in domestic NEED analysis.	• Domestic properties classified based on consumption for gas (less than 73,200kWh) and profile class for electricity (profiles 1 and 2 are domestic).
• Gas consumption between 100 kWh and 50,000kWh.	• Gas consumption below 73,200kWh.

• Electricity consumption between 100kWh and 25,000kWh.	• Electricity consumption below 50,000kWh and profile class 1 or 2.
• Data matched to other sources via the NLPG UPRN at property level.	• Data assigned to Lower Level Super Output Area[24].
• Suspected estimated readings removed.	

The differences lead to a small differences in mean consumption, but are important to provide confidence in the detailed analysis carried out with NEED. The comparisons with other sources confirm that the data in the NEED sample reflect the consumption for England as a whole.

[24] This means that for the Sub-national Consumption Statistics some properties can be assigned accurately if the street is identified even if the exact property is not know.

3. Homes Energy Efficiency Database

4.1 Introduction

HEED is a national database developed by the Energy Saving Trust (EST). It was set up to help monitor and target carbon reduction and fuel poverty work. It contains details of energy efficiency and micro-generation installations such as cavity wall insulation and solar hot water. It also includes information on the date each measure was installed.

HEED also includes data about property attributes (such as property age and type) and heating systems. However due to coverage and quality these data are not used in NEED.

4.2 Coverage

Data have been recorded in HEED since 1995 including activity reported from Government programmes, such as the Energy Efficiency Commitment (EEC) and Carbon Emissions Reduction Target (CERT), and activity reported by trade associations such as Gas Safe (formally CORGI) and FENSA.

Approximately 50 per cent of UK homes have a record in HEED. However there may not be full information for each of these records. For example, if a measure has been installed through a Government scheme then there may be information on the measure installed but no information on what other energy efficiency measures the property has, if they were not installed through a Government scheme. Table 4.1 shows how many records in the NEED sample had some kind of HEED record associated with it. It also sets out the number of measures recorded as being installed in properties in the NEED sample for each of the energy efficiency measures included in the analysis. These measures could have been installed in any year from 1995 to 2011.

Table 4.1: HEED data coverage in NEED sample

	Count	
	N	%
HEED Record	2,044,820	57%
Cavity wall insulation	640,890	18%
Loft insulation	536,410	15%
Heating measure	355,320	10%
Solid wall insulation[25]	1,670	0%

However, there is no information on measures that households have installed themselves (DIY measures) or measures installed at the time the property was built.

[25] Note that because of the small number of properties which have received solid wall insulation prior to 2010 the solid wall insulation sample used to estimate the impact of solid wall insulation has supplemented. All properties which were recorded as having had solid wall insulation at any time between 2005 and 2008 (and had valid gas consumption) were included in the analysis, not just those which were in the NEED sample used for all other analysis.

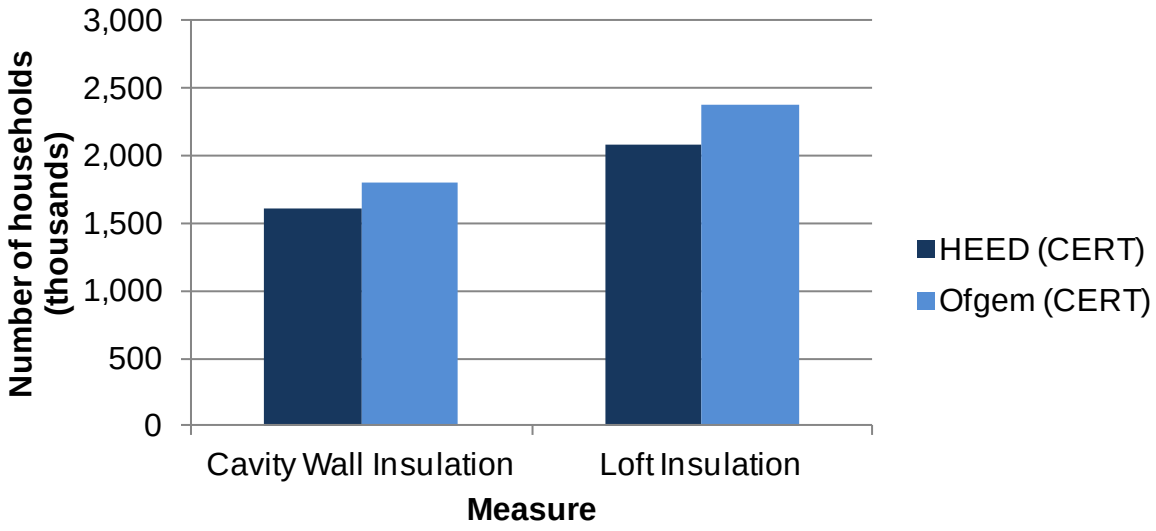
4.3 Data in HEED

For the majority of data used in NEED analysis, information is based on data EST receive from energy suppliers and covers measures installed through Government schemes (e.g. EEC and CERT). These data undergo validation before they are included in the HEED database. For example, checking that the same measure has not already been installed in the specified household.

Because the majority of measures recorded in HEED are measures installed through Government schemes, the types of households receiving measures are not representative of the population or housing stock as a whole. However, HEED does have good coverage of properties and which have received measures. More information on where measures are installed is provide in Annex D Section 3.

Figure 4.1 below shows the number of properties with cavity wall insulation and loft insulation installed through CERT up to September 2011. It shows that HEED includes a high proportion of the measures reported by suppliers to Ofgem. As no information is known about the specific properties receiving reported by Ofgem it is not possible to determine whether there is any bias in the HEED data, but the good coverage means that any bias should be small.

Figure 4.1: HEED and Ofgem reported CERT measures installed September 2011



It is expected that the gap between data reported by Ofgem and data included in HEED will reduce when energy suppliers provide the final property level information at the end of CERT.

Coverage of solid wall insulation is not as good as the cavity wall and loft insulation but should also improve as data is provided at the end of CERT. There will also be more solid wall insulation data available following the end of the Communities Energy Saving Programme (CESP), when data for this scheme should be included in HEED.

Coverage of boilers is also incomplete. Data come from a wider range of sources and therefore cover installations in a wider range of properties (see Annex D). However, there is currently no data available in HEED for installations since mid 2008. This is something EST are looking to rectify over the coming months.

When considering the quality of HEED data included in NEED is should also be noted that the installation dates associated with records are of varying quality, particularly for solid wall insulation, where it is not possible to distinguish when between 2005 and 2008 measures were installed.

4. Valuation Office Agency Data

5.1 Introduction

The Valuation Office Agency (VOA) is the central government agency responsible for valuing homes for council tax purposes[26]. The VOA has had responsibility for valuing properties for council tax since it was first introduced in 1993 and, before then, for the earlier system of domestic rates. Property attribute data was originally introduced in the 1970's in order to provide a simple system for understanding the main features and attributes of a property.

In order to maintain accurate and fair lists of council tax bandings, the VOA needs to keep the information it holds about properties up to date. It does this in a number of ways, including:

- Getting information from the local authority when a home is extended or altered to the extent that planning permission is required.
- Using voluntary questionnaires to enable the occupier to confirm information about a property.
- Other sources of freely available and publicly published information. For example, a contract with Calnea Analytics to access the Residata website which contains details of properties marketed through mouseprice.com since 2007.

In addition, the VOA will sometimes ask to visit a property when the information it needs cannot be ascertained from other sources. This can often be at the occupier's request; for example when they have challenged the council tax banding of their property and wish the VOA to carry out a review.

There are 16 individual property attributes collected, four of which are used in NEED analysis:

- Property type (detached, semi detached etc)
- Property age
- Floor area (m^2)
- Number of bedrooms

5.2 Coverage

The VOA Council Tax Database covers properties in England and Wales. The table below shows what proportion of properties are missing data for each of the variables used in this report. It shows the number of properties missing data for the VOA dataset as a whole (covering England and Wales) and for the sample of data used in NEED analysis.

[26] It does not set the level of council tax nor collect the money, which is the task of local government.

Table 5.1: VOA property attribute dataset missing data

	Property Age	Property Type	No. of Bedrooms	Floor Area
Missing - Full Dataset	1.0%	0.8%	1.5%	1.7%
Missing - NEED Sample	0.2%	0.0%	0.5%	0.7%

It shows that, for all variables, the coverage in the sample used for NEED is better than the coverage in the full dataset, and that for both versions of the dataset the coverage is very good. For the NEED sample, no variable is missing information for more than one per cent of records, with floor area the worst with 0.7 per cent of records missing this information.

The table below shows the categories of data used in the analysis for each of the VOA variables used. In most cases VOA have more detailed data than the categories shown; the VOA categories have been grouped to the categories set out for the purposes of the NEED analysis and presentation of results. Full details of the breakdowns included in the VOA dataset are available on its website[27].

Table 5.2: VOA property attribute data

	Property age	Property type	Number of bedrooms	Floor area (m^2)
Categories	Pre 1919	Detached	1	1-50
	1919-44	Semi detached	2	51-100
	1945-64	End terrace	3	101-150
	1965-82	Mid terrace	4	151-200
	1983-92	Bungalow	5+	201-250
	1993-99	Purpose built flat		Greater than 250
	Post 1999	Converted flat		

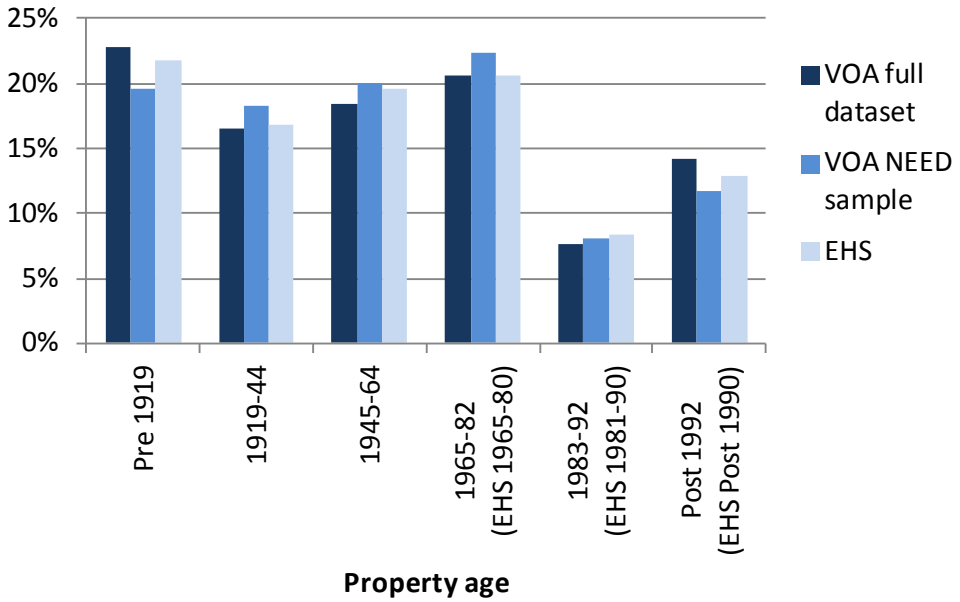
5.3 Summary of data and comparison with other sources

This section shows how the data in the NEED sample compares with property attribute data from the English Housing Survey (EHS)[28] . It also shows how the data in the NEED sample compare with the data for the property attribute dataset as a whole and the distribution of the population for each of these variables. When comparing these sources it should be noted that the VOA full dataset includes records for England and Wales while the other sources (EHS and NEED Sample) cover only England.

Figure 5.1 shows the proportion of properties in each age of construction band (property age) for each of the three sources of data.

[27] http://www.voa.gov.uk/corporate/Publications/DwellingHouseCodingGuide/index.html
[28] EHS data are from the English Housing Survey: Homes Report 2010

Figure 5.1: Comparison of sources – property age



The chart shows that there are some differences in the distributions, but overall the proportions are similar. There are two key reasons for the differences between the VOA full dataset and the dataset used in the NEED sample. The data in the VOA full dataset cover England and Wales, while the NEED sample only covers England and the EHS is based on a sample of households while the VOA is an administrative source with coverage of all properties.

Figure 5.2: Comparison of sources – property type

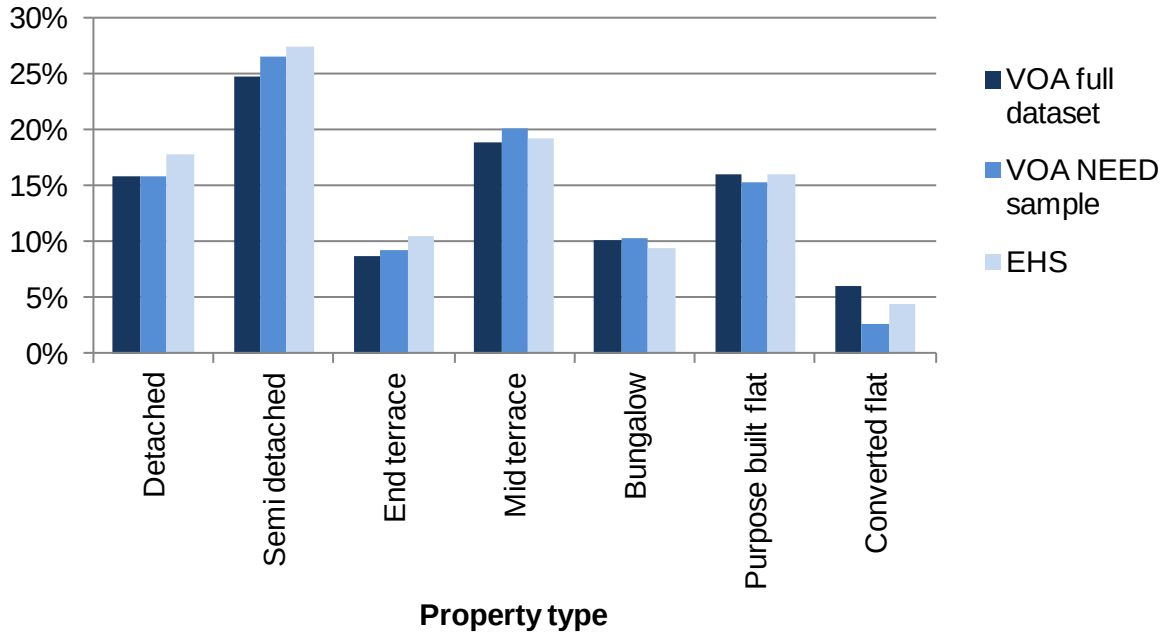


Figure 5.2 shows that in most cases the VOA full dataset and the data used for the NEED sample are similar. The biggest difference is for converted flats. This is due to matching issues with converted flats. As explained more fully in the methodology, flats are excluded from the analysis of impact of measures because of problems with matching converted flats in HEED.

Figure 5.3: Comparison of sources – number of bedrooms

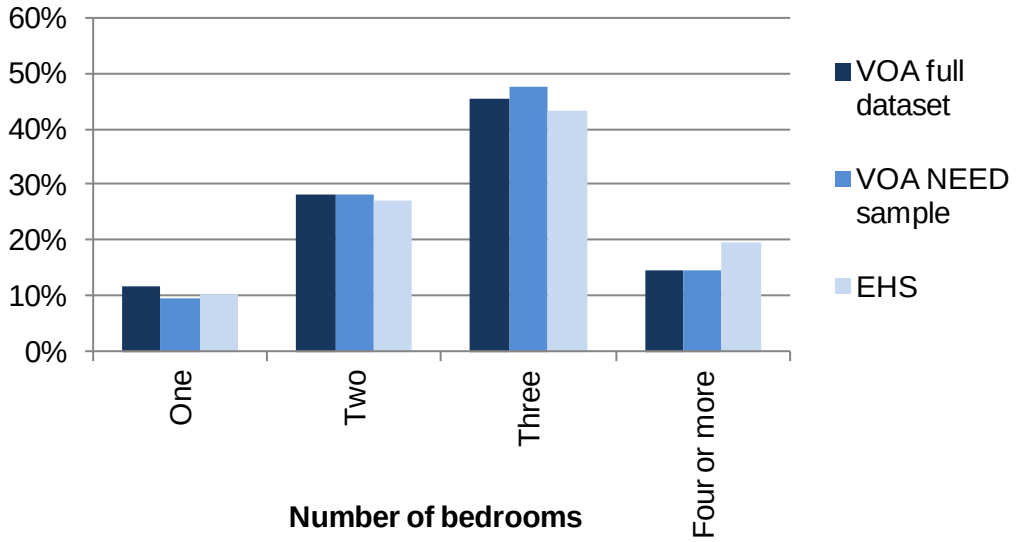
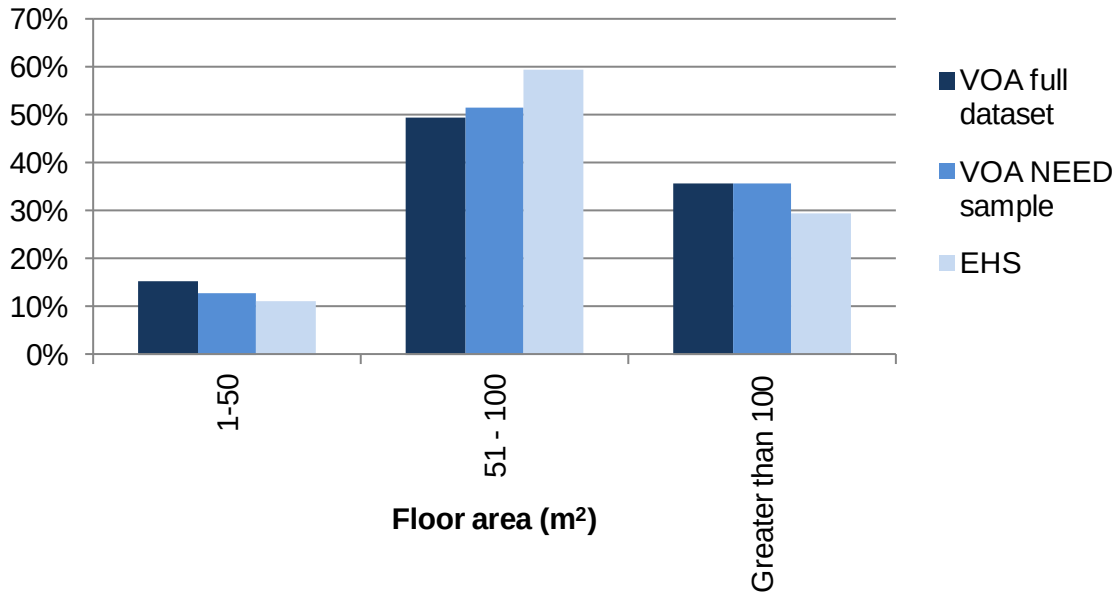


Figure 5.3 shows that the NEED sample has a very similar distribution to the VOA full database. There are some differences with the EHS data, particularly for properties with four or more bedrooms, but overall the sources are consistent. From the chart it is clear that the most dominant category is three bedroom properties with around 45 per cent of all properties having three bedrooms.

Due to the differences between the categories used for floor area on the EHS and those used in the NEED sample the floor area categories have been grouped for comparison, Figure 5.4.

Figure 5.4: Comparison of sources – floor area



It shows that the majority of properties have a floor area of between 51 and 100 square meters (approximately 550 to 1,100 square feet). It also shows that the VOA data are similar to that

recorded on the EHS[29]. The most significant cause for the difference is likely to be due to the difference in the definition of floor area between the two sources. The VOA use a different definition for houses and for flats. For houses the "Reduced Covered Area" is used while the "Effective Floor Area" is measured for flats. The reduced covered area is measured externally and is effectively the building's footprint. For flats, it is the internal floor area excluding some internal spaces such as bathrooms/showers and WCs, which are not excluded for houses. The EHS have a consistent definition for all properties[30]. There are also some small differences between the VOA full dataset and the sample used for NEED analysis. These are likely to be in part due to the differences between England and Wales and the small number of unmatched properties. This variable was not used in selecting the stratified sample, so there will also be some differences as a result of sampling.

5.4 Conclusion

The data in the VOA property attribute dataset have excellent coverage of properties in the UK and the comparisons with the EHS confirm that the distribution of data is consistent for all property attributes considered in the NEED analysis. Floor area shows the largest discrepancy between the EHS and NEED sample, but this is less than 10 per cent and is likely to result from the different definitions used for floor area between the two sources rather than an inaccuracy in either data source.

[29] Note the EHS reports 90-109m2 in the same category, for the purposes of comparison in the chart above half of this category has been included in the 51 – 100 category and half in the greater than 100 category.
[30] The variable floorx (usable floor area) is the internal floor area of the dwelling that excludes the area for stairwells and the area lost to partition walls e.g. a stair case is not considered as usable floor area because you cannot use that area to say put a cupboard there. The floor area of lofts are only included where the loft space is habitable with a fixed stairwell and the floor area if garages are only included where the garage is integral to the dwelling.

5. Experian Data

6.1 Introduction

Experian produce a dataset which models data for each property in the UK. DECC purchased a sample of these data for England for use in NEED. Data for each property are modelled by Experian based on other data sources including Experian surveys and aggregate published data. The data purchased by DECC are for 2009 and can be split into two groups:

- Property attribute data
- Household characteristics

The DECC sample is for approximately 3 million households, covering 81 per cent of the NEED analysis sample.

6.2 Property attribute data

The property attribute data purchased include residence type, property age and number of bedrooms. These variables are the same as variables available from the Valuation Office Agency and are purchased in order to allow analysis of data within DECC. These Experian variables are not used in the published report and therefore are not considered further in this section of the report.

6.3 Household characteristics

The household characteristic data purchased include:

- Household income
- Tenure
- Residency length
- Gender of household head
- Age of household head
- Number of adults
- Households with children

In this report only the income and tenure have been considered and therefore it is these variables which are covered in more detail below.

Household income

The household income variable identifies the likely household income for each property. The data are based on results from responses to Experian's consumer survey, which is then used alongside other predictive data (including Experian's person and household level demographics and Mosaic) to build a model.

Household income is available in ten income bands, table 6.1 below shows the bands and how many households are in each of these bands on the full Experian dataset. The full Experian dataset covers the UK, data purchased for NEED only cover a sample of England.

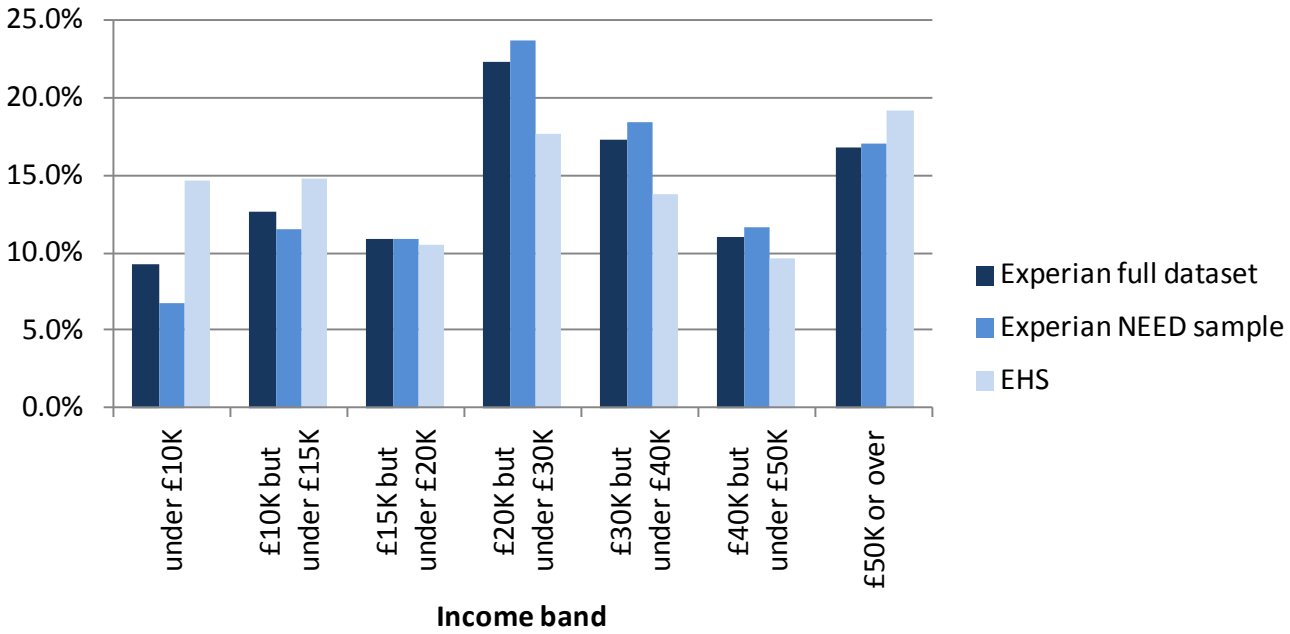
Table 6.1: Distribution of households by income band on the full Experian dataset

Band	Description	Households (%)
1	Less than £10,000	9.2 %
2	£10,000 - £14,999	12.6 %
3	£15,000-£19,999	10.9 %
4	£20,000 - £24,999	11.6 %
5	£25,000 - £29,999	10.7 %
6	£30,000 - £39,999	17.3 %
7	£40,000 - £49,999	11.0 %
8	£50,000 -£59,999	6.8 %
9	£60,000 - £74,999	5.7 %
10	£75,000 or more	4.3 %

It should be noted when interpreting any analysis of income in the NEED report that data for each property are modelled and therefore are indicative of the income a household is likely to have rather than an actual value for the current occupant of the property.

Experian have made an assessment of the quality of these data and conclude that on average, household income is accurate to £10,000. Based on Experian's assessment of the data, 32 per cent of properties are in the correct category and 54 per cent of properties are assigned to within 1 band of the correct category. Figure 6.1 shows how the distribution of income for the Experian dataset and the NEED sample compared with the income reported by the English Housing Survey (EHS). Note that some of the income categories from the Experian data have been grouped together to allow comparison with the categories used in the English Housing Survey.

Figure 6.1: Experian income data compared with EHS



The figure shows that Experian appears to be under assigning properties to the lower and highest income groups. This is consistent with DECC's understanding that the Experian income data is least reliable at the extremes. However, it should also be noted that the EHS is a survey and therefore subject to variation. Income is a self reported variable and therefore likely to be less reliable than the EHS variables considered in the previous section of this annex which are based on a physical survey of the property.

Tenure

Tenure data from Experian allocates each household in the UK to one of three categories; owner occupied, council/housing association or privately rented. The data are based on responses to Experian's lifestyle survey which are then used to predict the status of all properties. As with the household income variable, a model is used to predict the tenure for each property.

Table 6.2 shows the tenure variable values and how prevalent they are in the full Experian dataset.

Table 6.2: Distribution of households by tenure on the full Experian dataset

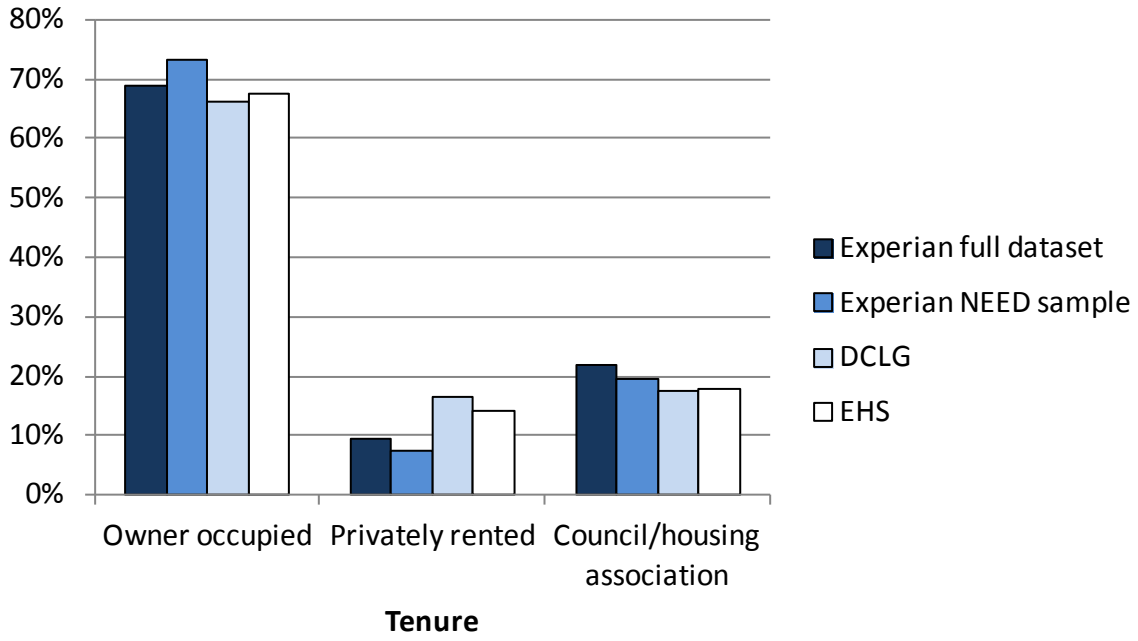
Band	Description	Households (%)
1	Owner occupied	68.9 %
2	Privately rented	9.4 %
3	Council/housing association	21.7 %

Experian's assessment of this variable suggests that 78.2 per cent of properties are allocated to the correct category. The accuracy of the variable varies within bands. For example 86 per cent of properties described as owner occupied in Experian's dataset are actually owner occupied, while only 43 per cent of properties allocated to privately rented are actually privately rented. For council/housing association housing the equivalent figure is 65 per cent.

Chart 6.2 shows how the Experian data compares with data from other sources at the national level[31].

[31] Note that the Experian full dataset covers the whole UK, while all other sources cover England only.

Chart 6.2: Experian tenure compared with other sources[32]



The chart shows that the number of properties assigned to each tenure category is similar for all sources. Differences between the Experian full dataset and the sample for NEED are likely to be due to the coverage and sample selection; the full Experian dataset covers all of the UK while the NEED sample only covers England. It appears that the Experian dataset as a whole and specifically the NEED sample allocates too many properties to the owner occupied category and too few to privately rented. The difference is most significant for the private rented sector where Experian data suggests approximately seven per cent of the NEED sample are owner occupied compared with around 15 per cent of households in England based on the DCLG and EHS datasets.

While the Experian data is valuable in order to provide an understanding of the properties in the NEED sample and how consumption and impacts of energy efficiency measures vary for different types of properties, it is important that interpretation of results relating to income and tenure is in the context of the limitations of the data.

[32] DCLG Table 104:
http://www.communities.gov.uk/housing/housingresearch/housingstatistics/housingstatisticsby/stockincludingvacants/livetables/
EHS Table 1.2 http://www.communities.gov.uk/documents/statistics/pdf/1937206.pdf

6. Conclusion

NEED is a valuable source of evidence on energy consumption and the impacts of energy efficiency measures, but it's value is dependent on the quality of the data used to form NEED.

This annex shows that in general the quality of data used in NEED are good, with excellent coverage of the population. In all cases, the distribution of data is broadly consistent with the other sources considered. At the property level, data from the administrative sources are more reliable than the data produced by Experian. Table 7.1 summarises the strengths and weaknesses of the data used in NEED.

Table 7.1: Strengths and weaknesses of data used in NEED

Data sources	Strengths	Weaknesses
Consumption data	• Covers Great Britain. • Good coverage of almost all properties (even post matching). • Data provided by energy suppliers. • Gas data are weather corrected.	• Based on billing data (sometimes estimated). • Gas and electricity years don't cover calendar year (or the same period as each other). • Domestic non-domestic split.
Homes Energy Efficiency Database (HEED)	• Has data for measures installed in homes in the UK including where measure is installed, and date of installation.	• Only covers measures installed through Government schemes; no information on measures installed by households themselves or installed when the property is built. • Matching of (converted) flats not reliable. • Only has a record for 57 per cent of properties in the NEED sample.
Valuation Office Agency (VOA)	• Covers every property in England and Wales. • Excellent coverage– data for each variable used is available for at least 99 per cent of properties in the NEED sample.	• No data for Scotland. • Some data may not be up to date.
Experian	• Data available for each household in the UK. • Best source of data at property level on household characteristics.	• Modelled data with variable accuracy at property level.

Overall, the data in NEED are of good quality. However, there are some weaknesses, and given the importance of the quality of the input data on the reliability of analysis, work will continue to be undertaken to improve the quality of data in NEED. This will include using Energy Performance Certificate (EPC) data to further validate some of the data in NEED and looking at alternatives to the Experian data.